Human Molecular Genetics, 2009, Vol. 18, Review Issue 1 **R60–R64**
doi:10.1093/hmg/ddp071

Lessons learnt from large-scale exon re-sequencing of the X chromosome

F. Lucy Raymond[1,*], Annabel Whibley[1], Michael R. Stratton[2] and Jozef Gecz[3,4]

[1]Cambridge Institute of Medical Research, University of Cambridge, Cambridge CB2 2XY, UK, [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK, [3]SA Pathology, Women's and Children's Hospital, North Adelaide, South Australia 5006, Australia and [4]Department of Paediatrics, The University of Adelaide, Adelaide, SA 5000, Australia

Received February 2, 2009; Revised and Accepted February 4, 2009

A candidate gene approach to identifying novel causes of disease is concept-limiting and in the new era of high throughput sequencing there is now no need to restrict the experiment to a few interesting genes. We have recently completed a large-scale exon re-sequencing project using Sanger sequencing technology to analyse approximately 1 Mb of coding sequence of the X chromosome in probands from >200 families with various forms of intellectual disability. We review the lessons learnt from this experience. Comparing large data sets will certainly reveal pathogenic mutations in genes that were not possible to identify previously. However, the task of distinguishing pathogenic mutations from rare sequence variants is not easy and is the most substantial challenge to the next decade. High-throughput technology has the attraction of being cheap, fast and comprehensive but for projects that require detailed coverage of a genomic region at an exhaustive level they may require a combination of large-scale with a small-scale follow-up of difficult regions to sequence. The number of rare truncating variants present in coding regions of the X chromosome that are not pathogenic was 1%. The importance of the quality of the starting material both clinically and molecularly and the number of sequence variants both rare and common that any one individual has across their coding sequence is discussed.

INTRODUCTION

We have recently published the results of a large-scale exon re-sequencing project to identify the cause of disease (1). This is the first of its kind where approximately 1 Mb of coding sequence of the X chromosome has been analysed in probands from >200 families with various forms of intellectual disability. While this is a considerable achievement with the current Sanger technology, this merely acts as a pilot study for the analyses of whole genomes in cohorts of individuals in the near future. The advent and ever increasing presence of the next-generation sequencing technology has fuelled an increase in large-scale sequencing projects and is heralding in the era of personalised genomes. Recently, the 1000 genome project has been launched (http://www.1000genomes.org) and in a few years the identification of the complete genomic sequence of an individual will be commonplace and will neither be limited by cost nor by technical expertise (2,3). We anticipate, however, that the interpretation of the data will become the central issue and concern, as it is not trivial. Having completed our study, it seems timely to review the lessons learnt from this experience, and to anticipate where the key challenges may lie in the future.

In our view, the overall strategy has been highly successful in the identification of novel X-linked genes that cause monogenic disease. This project has identified 12 new X-linked mental retardation (XLMR) genes that were not previously described and has also contributed to the discovery of four other genes (1,4–17). We also trust that several further XLMR genes are yet to be mined from the data. There is no doubt that this type of project is both feasible, achievable and a way forward to understand diseases that are genetically heterogenous (18,19).

*To whom correspondence should be addressed. Tel: +44 1223762609; Fax: +44 1223331206; Email: flr24@cam.ac.uk

DESIGN CHALLENGES

Epidemiological and genetic data for XLMR suggested that many genes were likely to cause the disease but individually the gene is a rare cause of disease although fully penetrant. While the positional and candidate gene-screening identified many XLMR genes, a new systematic approach was needed in order to identify the remainder (20).

As there were many potential disease genes to interrogate (>700 on the X chromosome), a rapid high-throughput gene abnormality detection methodology was needed. Also, the mutation detection method had to be accurate and not miss real but rare sequence variants (RSVs). For this reason, a direct polymerase chain reaction (PCR)-based DNA-sequencing method was chosen which used Sanger-based fluorescent sequence analysis rather than pre-screening samples with heteroduplex analysis or similar screening protocols. This particular project also gained from the hemizygous nature of the X chromosome in males such that mutation detection was always homozygous and relatively easy to identify.

In addition, large amounts of starting material was needed so that the availability of DNA was not rate-limiting. At the time we began, lymphoblastoid cell lines were created from most probands in the study so that high-quality DNA was available. DNA from low passage numbers was used to reduce the error rate from sequence alterations and chromosomal loss because of EBV transformation (21,22). Paired DNA samples from untransformed cells were always retained so that positive findings from the cell line DNA were always verified in the original DNA sample from blood. With the availability of accurate whole genome amplification protocols, it may no longer be necessary to create lymphoblastoid cell lines. However, the quality of the starting DNA material is critical. In our experience, old and poorly stored DNA samples often yield false-positive sequence abnormalities in high-throughput sequencing because of degradation of the DNA and whole genome amplification does not rectify this. In our study, several families were removed from further analysis as the number of missense mutations throughout the X chromosome were so high, >25, and were not reproducible in a second sample from the original stock. It was assumed that these were owing to poor quality starting DNA material. For the whole genome-amplified DNA, there was a close correlation between poor performance of the sample which we used, the age of the original sample and the amount of the remaining starting material.

As each gene was likely to be rare, we ensured that the hit rate for any one gene was maximized by starting with DNA from a large number of families with good evidence of X-linkage. As the disease is so highly heterogeneous, pooling of sample data was not feasible and thus by its very nature, this project required a large international collaboration between clinicians and molecular geneticists (the IGOLD project; http://goldstudy.cimr.cam.ac.uk/). This has been invaluable and has ensured that the sample selected for sequencing was derived from a larger cohort that could be refined. Families were selected out of a possible cohort of >600 families, where the evidence of an X-linked monogenic disease was greatest, there were high-quality DNA samples available and where a pathogenic mutation in a known MR gene had been excluded or the clinical picture was not typical of a syndrome where the causative gene had already been determined, e.g Coffin Lowry syndrome. All subjects were pre-screened clinically by an experienced clinical geneticist, a recent karyotype analysis was normal at a 500G level and Fragile X syndrome was excluded molecularly. This meant that the clinical data were of high quality and the samples were enriched for families, where a novel disease-causing gene was likely to be present but it created ascertainment biases that precluded any meaningful prevalence assessment of any new gene when it emerged in the context of XLMR as a whole. The strategy did, however, enable several new syndromes to be defined from the cohort once we had obtained the sequencing results (*CUL4B*, *SLC9A6*, *MED12*).

Careful assessment of the pathogenic significance of a sequence variant in this project has been a primary aim as the degree of coding sequence variation over a large number of genes from any one individual was not known. In order to do this, having sufficient samples from appropriate controls was important and in some cases over 1000 controls were analysed to identify the frequency of RSVs in each gene. Also, we used detailed segregation analysis in families to assess a putative pathogenic sequence variant. This enabled a number of sequence variants to be definitively excluded because of non-segregation of the variant with disease.

TECHNICAL CHALLENGES

One of the major advantages of high-throughput methodologies is their unbiased nature. The field of XLMR gene discovery was previously subject to various biases including small sets of families and the use of candidate gene selection criteria. This project aimed to interrogate all or as many as possible of the 720 VEGA-annotated genes on the X chromosome using unique sequence primer design. Only regions of intra- or inter-chromosomal similarity were not interrogated, which was only a significant problem for the X-Y pseudoautosomal region of the X.

In addition, a high-throughput method requires a simple experimental protocol in order to maximize speed and efficiency. This means that both automated primer design and relatively inflexible PCR protocols are required with the inevitable associated reduction of coverage for each gene compared with gene-specific protocols. Although the data set reported on the families in the cohort included sequence analysis of $\geq$75% of all exons on the X chromosome, the high-throughput analysis is not necessarily complete for all genes. The coverage was biased according to GC content, such that exon 1 was relatively poorly represented in all genes and some genes with a high overall GC content, e.g. *ARX* were particularly poorly sequenced.

We used high-throughput sequence analysis in order to detect new genes that cause disease as we predicted that the greatest number of disease-causing mutations were small sequence variants that could be detected by this method. In principle, this methodology would technically also detect whole exon deletions because of failure to amplify a PCR product, but in practice this was not analysed extensively, as

the false-positive rate owing to technical errors in the PCR amplification was frequently the cause of failure to amplify rather the presence of a deletion. On completion of the sequence analysis we found that 16% of the cohort did not have any unique missense variant detected at all. As many of the families with no significant sequence variant detected are multigenerational families with a high likelihood of having a single pathogenic mutation, additional methods such as array CGH are now being used to systematically identify whole exon deletions and duplications as a cause of disease.

Data handling is a major issue in this type of project, manual inspection of all the traces was not possible as >3 million traces were generated. However, the commercially available sequence trace analysis packages were not sufficiently robust to interpret the abnormal traces they detected and a degree of manual inspection was needed to assess traces where sequence variants were detected in both sequencing directions (23). The data analysis required the expertise of a genome centre (http://www.sanger.ac.uk) where in-house bioinformaticians could develop the analysis programmes and where there was sufficient capacity to store these large data sets, and also, where people were available to be deployed to inspect the abnormal sequence traces for a period of time in the project.

In large cohort studies the potential to generate huge amounts of data is unquestionable. The issue of how this should be made available to the scientific community is a little more complex. For families who took part in our study, the commonest reasons for taking part was their desire to know what the genetic cause of the disease was in their family, they wanted to know if others in the family were at risk of having children with a similar condition and also, they wanted to contribute to the overall knowledge of learning disability to improve diagnoses for families in general (24). In this study, the data set for each family effectively provides a series of unique identifiers for each patient on the X chromosome. This creates a tension between the need for the scientific community to have access to large scale data sets of sequence and the need to preserve confidentiality of all patients taking part in a study. This is a relatively new situation and while participants consented to take part in the study they did not explicitly consent for their personal genotype to be placed on the internet for all to see. We have thus limited access to the data of researchers who participate in the general academic peer review publication process as a means of ensuring that the data set is used for scientific and medical benefit and thus conforms to the consent but we also hope that the data will be used by the scientific community as much as possible in order to support the aims of publicly funded research (http://goldstudy.cimr.cam.ac.uk/).

INTERPRETIVE CHALLENGES

Assigning a gene to a disease is relatively straightforward where the phenotype is clear and the disease is predominantly because of truncating mutations in a single gene. This proved to be the case for *FRMD7*, where most cases of X-linked congenital nystagmus were owing to truncating mutations in the single novel gene we identified (5). However, for clinically more heterogenous conditions such as XLMR the criteria used to judge whether or not a sequence variant is the cause of disease needs more careful analysis. For recessive diseases, the identification of a truncating or loss of function variant is likely to cause disease, however, in genes where there is no prior evidence of variants causing disease, the presence of a truncating variant in a single family where the disease is heterogenous is probably not sufficient. We identified seven XLMR genes, where we were able to detect multiple truncating variants in the gene and yet found no truncating variants in a large control set that were fully sequenced for the gene in question. However, we identified a further 19 genes where a single truncating variant was identified in the disease group. Many of these rare variants were then excluded by segregation analysis or more extensive analysis of a control population. Without the extended pedigrees we would not have been able to exclude so many of these genes. This suggests that up to 1% of the genes on the X chromosome can sustain loss-of-function variants and they do not appear to have a detrimental effect. It is likely that a similar proportion of genes on autosomes are potentially redundant although the proportion of redundant genes on the X chromosome may be greater than for autosomes (25–27). We believe that the identification of a single rare deletion of a gene or part of a gene or a truncating variant is not sufficient evidence that the gene abnormality is pathogenic and without the assistance of parental samples and extended pedigrees this can be difficult to interpret.

Because the sequencing analysis covered all coding exons on the X chromosome, this project acts as a screen shot of the human genome at a specific moment in the evolutionary continuum of the X chromosome. Provided there were defined exons on a genome browser an amplicon was investigated. The genome browsers make little distinctions between coding and non-coding parts of exons, nor pseudogenes or transposons or genome sequences that are travelling between these states by natural selection (25–28). The significance of a sequence variant will depend on the location of the variant within the sequence, the background polymorphism rate of the gene and the functional redundancy of the protein in the organism. Although prediction programmes like Polyphen, Panther and SIFT, and measures of protein evolution are useful in assessing whether a variant is likely to be pathogenic, the current tools available to determine pathogenicity are limited (29,30).

In the first analysis of the data we were surprised not to identify more truncating mutations in genes than we did (38/208). As most of the XLMR genes that had been identified prior to the project starting contained truncating mutations as the predominant disease-causing class of mutation, we assumed that this would continue in the cohort we analysed looking for novel genes that cause mental retardation. There are, however, precedents for genes on the X chromosome where truncating mutations are likely to be lethal such as *MECP2*, *NEMO*, *CDKL5* and more recently *CASK* and this may explain why very few families were explained by a truncating mutation (1,31–34). It also seems possible that the previous XLMR studies were biased towards identifying the protein truncating variants and only the difficulty to find

genes remained. It is also possible that a significant number of truncating variants were missed by the high-throughput strategy as only 75% of the X chromosome was covered on average per family and thus 25% of mutations would have been missed.

Perhaps, the most surprising feature of the data set is the number of rare coding sequence variants that were identified in any one individual. The project was focused on the coding sequence of the X chromosome, an area of the genome, where fewer sequence variants are found compared with non-coding regions. Nevertheless, for any one individual we identified on average three to four unique amino acid sequence changes over and above the expected high number of common non-synonymous and synonymous coding polymorphisms. Perhaps, it was naïve to imagine that for each family we studied with mental retardation, a single pathogenic variant would be identified as the cause of the disease as on average >5000 exons were sequenced in any one individual. Of the unique sequence variants we identified, the majority are likely not to be the cause of the mental retardation phenotype in the patients as the pedigree structure suggests that the disease is likely to be owing to a single gene defect. Technically, the other unique sequence variants found in an individual are not classed as polymorphisms as they are less frequent than 1% of the population. We have termed these RSVs. The problem is the difficulty in establishing in any one individual the sequence changes that are pathogenic and cause mental retardation and which are RSVs of no pathogenic significance or of significance to the individual but not contributing to the MR phenotype. For some, assigning pathogenicity may be easier as the variant occurs in highly conserved residues in an active site of a protein or in a protein known to be involved in syndromic mental retardation or it can be dismissed as the RSV does not segregate with disease in the family. However, where two RSVs are in linkage disequilibrium or within neighbouring genes it is unlikely that segregation analysis will resolve the issue of pathogenicity. Currently, there are few resources available to identify whether the rare variants have been identified elsewhere before. The HapMap project (http://www.hapmap.org/) and other SNP identification resources such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP), all concentrate on the common alleles in the population and do not report RSVs routinely. For many of the RSVs, we identified that a bioinformatic approach alone is insufficient to determine pathogenicity and a detailed assessment of the functional consequence of each sequence variant is needed. In practice, no high throughput or even medium throughput methodology has been developed for this and this is now the rate-limiting step in the identification of further genes that cause disease.

SUMMARY

A candidate gene approach to identifying novel causes of disease is concept-limiting and in the new era of next generation sequencing there is no need to restrict the experiment to a few interesting genes. We are entering a data-rich period where more and more sequence analysis is possible. Comparing large data sets will reveal genetic pathology that was not possible to identify previously. However, the task of distinguishing pathogenic mutations from RSVs is not resolved and is perhaps the most substantial challenge to the next decade. High-throughput technology has the attraction of being new and big but the failing of lack of adaptability means that projects that require detailed coverage of a candidate region at an exhaustive level may require a combination of large-scale with a small-scale follow-up of difficult regions to sequence. The number of rare truncating variants present in the genome that are not pathogenic was a surprise. Until we came to the final analysis we had not appreciated the importance of the quality of the starting material both clinically and molecularly, how to release the data to maintain the best interests of all parties involved and the sheer number of sequence variants any one individual has across their coding sequence.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the many families with members with mental retardation or learning disability who agreed to participate in the studies performed by the IGOLD consortium. Also to P. Tarpey, R. Smith, A. Futreal, C. Schwartz, R. Stevenson, G. Turner, A. Hackett and M. Field who have all made major contributions to the IGOLD consortium.

Conflict of Interest statement. None declared.

FUNDING

The Wellcome Trust, Action Medical Research, NIHR, the New South Wales Department of Health and the Australian NHMRC.

REFERENCES

1. Tarpey, P., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Tofts, C., Dicks, E., Menzies, A. *et al.* (2009) A systematic, large scale resequencing screen of the X chromosome coding exons in mental retardation. *Nat. Genet.*, in press.
2. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
3. Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.
4. Tarpey, P., Parnau, J., Blow, M., Woffendin, H., Bignell, G., Cox, C., Cox, J., Davies, H., Edkins, S., Holden, S. *et al.* (2004) Mutations in the *DLG3* gene cause nonsyndromic X-linked mental retardation. *Am. J. Hum. Genet.*, **75**, 218–324.
5. Tarpey, P., Thomas, S., Sarvananthan, N., Mallya, U., Lisgo, S., Talbot, C.J., Roberts, E.O., Awan, M., Surendran, M., McLean, R.J. *et al.* (2006) Mutations in FRMD7, a newly identified member of the FERM family, cause X-linked idiopathic congenital nystagmus. *Nat. Genet.*, **38**, 1242–1244.
6. Tarpey, P.S., Raymond, F.L., Nguyen, L.S., Rodriguez, J., Hackett, A., Vandeleur, L., Smith, R., Shoubridge, C., Edkins, S., Stevens, C. *et al.* (2007) Mutations in UPF3B, a member of the nonsense-mediated mRNA decay complex, cause syndromic and nonsyndromic mental retardation. *Nat. Genet.*, **39**, 1127–1133.
7. Tarpey, P.S., Raymond, F.L., O'Meara, S., Edkins, S., Teague, J., Butler, A., Dicks, E., Stevens, C., Tofts, C., Avis, T. *et al.* (2007) Mutations in CUL4B, which encodes a ubiquitin E3 ligase subunit, cause an X-linked mental retardation syndrome associated with aggressive outbursts,

seizures, relative macrocephaly, central obesity, hypogonadism, pes cavus, and tremor. *Am. J. Hum. Genet.*, **80**, 345–352.
8. Tarpey, P.S., Stevens, C., Teague, J., Edkins, S., O'Meara, S., Avis, T., Barthorpe, S., Buck, G., Butler, A., Cole, J. *et al.* (2006) Mutations in the gene encoding the Sigma 2 subunit of the adaptor protein 1 complex, AP1S2, cause X-linked mental retardation. *Am. J. Hum. Genet.*, **79**, 1119–1124.
9. Raymond, F.L., Tarpey, P.S., Edkins, S., Tofts, C., O'Meara, S., Teague, J., Butler, A., Stevens, C., Barthorpe, S., Buck, G. *et al.* (2007) Mutations in ZDHHC9, which encodes a palmitoyltransferase of NRAS and HRAS, cause X-linked mental retardation associated with a Marfanoid habitus. *Am. J. Hum. Genet.*, **80**, 982–987.
10. Field, M., Tarpey, P., Boyle, J., Edkins, S., Goodship, J., Luo, Y., Moon, J., Teague, J., Stratton, M.R., Futreal, P.A. *et al.* (2006) Mutations in the *RSK2* (RPS6KA3) gene cause Coffin-Lowry syndrome and nonsyndromic X-linked mental retardation. *Clin. Genet.*, **70**, 509–515.
11. Field, M., Tarpey, P.S., Smith, R., Edkins, S., O'Meara, S., Stevens, C., Tofts, C., Teague, J., Butler, A., Dicks, E. *et al.* (2007) Mutations in the *BRWD3* gene cause X-linked mental retardation associated with macrocephaly. *Am. J. Hum. Genet.*, **81**, 367–374.
12. Schwartz, C.E., Tarpey, P.S., Lubs, H.A., Verloes, A., May, M.M., Risheg, H., Friez, M.J., Futreal, P.A., Edkins, S., Teague, J. *et al.* (2007) The original Lujan syndrome family has a novel missense mutation (p.N1007S) in the *MED12* gene. *J. Med. Genet.*, **44**, 472–477.
13. Molinari, F., Foulquier, F., Tarpey, P.S., Morelle, W., Boissel, S., Teague, J., Edkins, S., Futreal, P.A., Stratton, M.R., Turner, G. *et al.* (2008) Oligosaccharyltransferase-subunit mutations in nonsyndromic mental retardation. *Am. J. Hum. Genet.*, **82**, 1150–1157.
14. Gilfillan, G.D., Selmer, K.K., Roxrud, I., Smith, R., Kyllerman, M., Eiklid, K., Kroken, M., Mattingsdal, M., Egeland, T., Stenmark, H. *et al.* (2008) SLC9A6 mutations cause X-linked mental retardation, microcephaly, epilepsy, and ataxia, a phenotype mimicking Angelman syndrome. *Am. J. Hum. Genet.*, **82**, 1003–1010.
15. Wu, Y., Arai, A.C., Rumbaugh, G., Srivastava, A.K., Turner, G., Hayashi, T., Suzuki, E., Jiang, Y., Zhang, L., Rodriguez, J. *et al.* (2007) Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated with moderate cognitive impairment in humans. *Proc. Natl Acad. Sci. USA*, **104**, 18163–18168.
16. Froyen, G., Corbett, M., Vandewalle, J., Jarvela, I., Lawrence, O., Meldrum, C., Bauters, M., Govaerts, K., Vandeleur, L., Van Esch, H. *et al.* (2008) Submicroscopic duplications of the hydroxysteroid dehydrogenase HSD17B10 and the E3 ubiquitin ligase HUWE1 are associated with mental retardation. *Am. J. Hum. Genet.*, **82**, 432–443.
17. Dibbens, L.M., Tarpey, P.S., Hynes, K., Bayly, M.A., Scheffer, I.E., Smith, R., Bomar, J., Sutton, E., Vandeleur, L., Shoubridge, C. *et al.* (2008) X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment. *Nat. Genet.*, **40**, 776–781.
18. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S. *et al.* (2007) Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.*, **80**, 779–791.
19. Jensen, L.R., Lenzner, S., Moser, B., Freude, K., Tzschach, A., Wei, C., Fryns, J.P., Chelly, J., Turner, G., Moraine, C. *et al.* (2007) X-linked mental retardation: a comprehensive molecular screen of 47 candidate genes from a 7.4 Mb interval in Xp11. *Eur. J. Hum. Genet.*, **15**, 68–75.
20. Ropers, H.H., Hoeltzenbein, M., Kalscheuer, V., Yntema, H., Hamel, B., Fryns, J.P., Chelly, J., Partington, M., Gecz, J. and Moraine, C. (2003) Nonsyndromic X-linked mental retardation: where are the missing mutations? *Trends Genet.*, **19**, 316–320.
21. Gao, Y., Lu, Y.J., Xue, S.A., Chen, H., Wedderburn, N. and Griffin, B.E. (2002) Hypothesis: a novel route for immortalization of epithelial cells by Epstein-Barr virus. *Oncogene*, **21**, 825–835.
22. Kamranvar, S.A., Gruhne, B., Szeles, A. and Masucci, M.G. (2007) Epstein-Barr virus promotes genomic instability in Burkitt's lymphoma. *Oncogene*, **26**, 5115–5123.
23. Dicks, E., Teague, J.W., Stephens, P., Raine, K., Yates, A., Mattocks, C., Tarpey, P., Butler, A., Menzies, A., Richardson, D. *et al.* (2007) AutoCSA, an algorithm for high throughput DNA sequence variant detection in cancer genomes. *Bioinformatics*, **23**, 1689–1691.
24. Ponder, M., Statham, H., Hallowell, N., Moon, J.A., Richards, M. and Raymond, F.L. (2008) Genetic research on rare familial disorders: consent and the blurred boundaries between clinical service and research. *J. Med. Ethics*, **34**, 690–694.
25. Emerson, J.J., Kaessmann, H., Betran, E. and Long, M. (2004) Extensive gene traffic on the mammalian X chromosome. *Science*, **303**, 537–540.
26. Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jegou, B. and Kaessmann, H. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.*, **6**, e80.
27. Vinckenbosch, N., Dupanloup, I. and Kaessmann, H. (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA*, **103**, 3220–3225.
28. Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.*, **3**, e357.
29. Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
30. Valdar, W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
31. Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.*, **23**, 185–188.
32. Kenwrick, S., Woffendin, H., Jakins, T., Shuttleworth, S.G., Mayer, E., Greenhalgh, L., Whittaker, J., Rugolotto, S., Bardaro, T., Esposito, T. *et al.* (2001) Survival of male patients with incontinentia pigmenti carrying a lethal mutation can be explained by somatic mosaicism or Klinefelter syndrome. *Am. J. Hum. Genet.*, **69**, 1210–1217.
33. Najm, J., Horn, D., Wimplinger, I., Golden, J.A., Chizhikov, V.V., Sudi, J., Christian, S.L., Ullmann, R., Kuechler, A., Haas, C.A. *et al.* (2008) Mutations of CASK cause an X-linked brain malformation phenotype with microcephaly and hypoplasia of the brainstem and cerebellum. *Nat. Genet.*, **40**, 1065–1067.
34. Weaving, L.S., Christodoulou, J., Williamson, S.L., Friend, K.L., McKenzie, O.L., Archer, H., Evans, J., Clarke, A., Pelka, G.J., Tam, P.P. *et al.* (2004) Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. *Am. J. Hum. Genet.*, **75**, 1079–1093.